Nan Tang

Senior Scientist, Qatar Center for Artificial Intelligence
Qatar Computing Research Institute (QCRI), HBKU, Qatar Foundation
Doha, Qatar

+974 66700540
+974 44542850
ntang@hbku.equ.qa
da.qcri.org/ntang/
DOB: 28 September 1980

Research Interest (Data Preparation for ALL)

Traditional DP	Data preparation (DP) theory and systems, especially trusted and explainable solutions
DL for DP	Deep learning (DL) for DP, especially models for human-easy but computer-hard DP tasks
DP for ML	DP for data-centric machine learning (ML), especially for co-debugging data and label problems
VIS for DP	Visualization (VIS) for DP – natural language-based VIS recommendation and generation

Education

2004/07–2007/12 **Ph.D.**, Systems Engineering & Engineering Management
The Chinese University of Hong Kong, Hong Kong
• Thesis: Efficient XPath Query Processing in Native XML Databases
• Co-Supervisors: Jeffrey Xu Yu and Kam-Fai Wong

2001/09–2004/01 **M.Sc.**, Computer Science
Northeastern University, China
• Thesis: Parallel XML Databases
• Supervisor: Guoren Wang

1997/09–2001/01 **B.S.**, Computer Science, *Northeastern University, China*

Professional Experience

2015/04–now *Senior Scientist*, Qatar Center for Artificial Intelligence, `QCRI`, Qatar.

2017/07–08 *Visiting Scientist*, `MIT`, US. Worked on the Data Civilizer project, with Michael Stonebraker, Samuel Madden, and Armando Solar-Lezama.

2011/12–2015/03 *Scientist*, Data Analytics, `QCRI`, Qatar.

2010/02-2012/01 *Research Fellow*, `University of Edinburgh`, UK. Worked on data cleaning and graph algorithms, with Wenfei Fan.

2008/02-2010/01 *Scientific Staff Member*, `CWI` (the national research institute for mathematics and computer science), the Netherlands. Worked on column-store database MonetDB and distributed XQuery processing, with Peter Boncz.

2007/03–08 *Visiting Scholar*, `University of Waterloo`, Canada. Worked on XML indexing and query rewriting, with Tamer Özsu.

Awards

2018 *SIGMOD 2020 Reproducibility Award*: "Raha: A Configuration-Free Error Detection System".

2018 *Best papers of ICDE 2018*: "Discovering Mis-Categorized Entities".

2015 *Best papers of VLDB 2015*: "Lightning Fast and Space Efficient Inequality Joins".

2012 *Best papers of ICDE 2012*: "Incremental Detection of Inconsistencies in Distributed Data".

2010 *The Best Paper Award of VLDB 2010*: "Towards Certain Fixes with Editing Rules and Master Data".

2009 *Best papers of ICDE 2009*: "Projective Distribution of Full-Fledged XQuery".

Research at QCRI (Dec 2011–Present)

Traditional DP

- **Error detection.** I have collected many real-world data errors (PVLDB'16), and devised methods to detect different types of data errors (ICDE'18, KDD'18, SIGMOD'17, SIGMOD'19).
- **Trusted data repairing.** In practice, users are hesitant to see their data being automatically repaired, unless these repairs are ensured to be correct or explainable. My works on trusted data repairing include rule-based methods (SIGMOD'14, PVLDB'20), using master data (ICDE'15), knowledge bases (SIGMOD'15, ICDE'17), and human-in-the-loop (SIGMOD'16).
- **A commodity data cleaning system.** There was no commodity platform similar to general purpose DBMSs that can be easily customized and deployed to solve application-specific data quality problems. I led the project NADEEF (SIGMOD'13, SIGMOD'15), which provides a unified programing interface for *declaratively* specifying what are data errors and (possibly) how to fix them, and a core that holistically handles the detection and repairing of data errors. NADEEF has five filed US/EU patents, and we made a startup attempt.
- **DP as a Service.** Collaborated with MIT, we are building DATA CIVILIZER (CIDR'17) with a suite of prebuilt tools for solving end-to-end data preparation problems. During interacting with real-world scenarios, we also developed the following new components.
  - *Data discovery*: we have solved the problem of linking the datasets in a data lake for data discovery (ICDE'18), driven by real-world scenarios from Merck and Scotiabank.
  - *Interpretable entity resolution*: we propose to use *program synthesis by examples* to discover rules on entity matching (PVLDB'18) and entity consolidation (ICDE'19), driven by TAMR use cases that their customers require their entity resolution solutions to be interpretable.
  - *Relational table storage and query co-optimization:* we develop deductive program synthesis algorithms for co-optimizing data storage and query plan (OOPSLA'20), to improve the efficiency of pre-defined workflows with relatively static data and parameterized queries.
  - *Data debugging*: we study the problem of debugging data (not code) problems (CIDR'20) in data science pipelines, by working with Massachusetts General Hospital (MGH), Intel MIT lab, and All Chicago (an organization that helps homeless people in Chicago).
- **Collaborative DP.** In DP, oftentimes, human-in-the-loop is not enough. We need crowd-in-the-loop to collaboratively clean and annotate data. Collaborated with **University of Wisconsin-Madison**, we are building such systems.
  - *CoClean* [SIGMOD Demo'20]: We built an Overleaf-like platform, on top of Python Pandas DataFrame, that enables multiple users to collaboratively clean the same dataset, which handles user synchronization and annotation aggregation, and allows customization.

DL for DP

Despite all the theoretical and systematic efforts, DP still dominates data scientists' time. My ***goal*** is to automate human-easy but computer-hard tasks with pre-trained DL models.

- **Relational pre-trained transformer (RPT).** I designed RPT (PVLDB'21), a denoising autoencoder for *tuple-to-X* models ("*X*" could be tuple, token, label, JSON, and so on) that supports a wide range of data preparation tasks such as data cleaning, auto-completion, schema matching, entity resolution, value normalization, data transformation, data annotation, information extraction, and so forth. RPT is pre-trained for a *tuple-to-tuple* model with fill-in-the-blank style denoising objectives, and can be fine-tuned for multiple tasks.
- **Entity resolution.** We study to improve blocking and entity matching by DL such as semi-supervised representation learning and contrastive learning (PVLDB'18).

DP for ML

- **Adaptive data augmentation:** We use Generative Adversarial Networks (GANs) for adapting the training data to the testing data for enhancing supervised ML, when the training and the testing data have different missing values patterns (PVLDB'21).

VIS for DP

My ***goal*** is to help users understand the tables discovered from a data lake beyond eyeballing, in particular for the DATA CIVILIZER project. I did this line of research by co-supervising a Ph.D. student from **Tsinghua University**, China.

- **Automatic visualization:** We developed DEEPEYE (ICDE'18) that trains a binary classifier for deciding good/bad visualizations, and a ranking model to rank good visualizations. This technique has been used by Tencent and ByteDance.
- **Natural language to visualization (NL2VIS) benchmark:** We produced the first NL2VIS benchmark, called NVBENCH (SIGMOD'21), with the goal to advance the field of NL2VIS. We are the first to apply neural translation model on NL2VIS and demonstrate that it significantly outperforms the state-of-the-art NLP semantic parser-based approaches.
- **COVID-19 dashboards.** I developed and led the following dashboards.
  - *Qatar situation dashboard:* used by MOI Qatar; showcased on AlJazeera, Turkish TV.
  - *COVID mobility analysis:* used by MOPH Qatar, Kuwait Health Ministry, and Nigerian National Bureau of Statistics.
  - *COVID data and mobility analysis in China* [IEEE Data Eng. Bull.'20]. In early 2020, we built a COVID-19 dashboard that attracted millions of visits, and worked with China Mobile to visualize and analyze the trajectories of infected persons in Beijing.

Research at University of Edinburgh (Feb 2010–Dec 2011)

Data Cleaning

Worked on data cleaning using master data (PVLDB'10, the Best Paper Award), interacting different types of data quality rules (SIGMOD'11), incrementally detecting errors in distributed data (ICDE'12), and inferring data currency and consistency for conflict resolution (ICDE'13).

Graph Algorithms

Worked on graph pattern matching algorithms (PVLDB'10) and then added regular expressions to graph pattern queries (ICDE'11).

Research at CWI (Feb 2008–Jan 2010)

MonetDB

Worked on efficiently supporting updates in column-stores using packed memory arrays. building space-economical Q-gram index for exact string matching over a 400+GB dataset (CIKM'09), and enabling efficient distribution of full-fledged XQuery on top of MonetDB/XQuery (ICDE'09), for supporting use cases in the Netherlands Forensic Institute.

Teaching and Mentoring Experience

Mentored Interns and Postdocs, QCRI

Hakim Qahtan	Ph.D., KAUST, Saudi Arabia (now assistant professor at Utrecht)	2017/09-2020/08
Jinsong Guo	Ph.D., University of Oxford, UK	2017/03-2017/09
Sibo Wang	Ph.D., NTU, Singapore (now assistant professor at CUHK)	2016/06-2016/11
Dong Deng	Ph.D., Tsinghua University, China (now assistant professor at Rutgers)	2016/06-2016/08
Sourav Medya	Ph.D., UC Santa Barbara, US	2016/06-2016/08
Qing Chen	Master, Fudan University, China (now Ph.D. at Zurich University)	2015/07-2016/04
Jian He	Master, Tsinghua University, China (now at Google)	2014/11-2015/02
Matteo Interlandi	Ph.D., University of Modena, Italy (now at MSR)	2014/03-2014/05
Chu Xu	Ph.D., U. of Waterloo, Canada (now assistant professor at Georgia Tech)	2013/05-2014/07
Jiannan Wang	Ph.D., Tsinghua University, China (now associate professor at Simon Fraser)	2012/12-2013/02
Yu Tang	Master, Hong Kong University, HK (now Ph.D. at Oxford)	2012/11-2013/01
Amr Ebaid	Ph.D., Purdue University, US (now at Google)	2012/04-2013/01
Ahmed Eldawy	Ph.D., U. of Minnesota, US (now assistant professor at UC Riverside)	2012/01-2012/05
Michele Dallachiesa	Ph.D., University of Trento, Italy	2012/01-2012/05

External Ph.D. Advisor, Tsinghua University

Yuyu Luo	Ph.D., Tsinghua University, China (topic: VIS for DP)	2018/09-present

Teaching, University of Edinburgh, UK (Tutorials)

Applied Databases 2010/09-11

Teaching, The Chinese University of Hong Kong, Hong Kong (Tutorials)

Digital Logical and Systems 2006/09-12, 2007/09-12

Fundamentals of Information Systems 2004/09-12, 2006/01-05

Information Systems Design & Analysis 2005/01-05

Selected Professional Activities and Services

PC Member SIGMOD Exhibition Chair (2021), SIGMOD (2015, 2017–2020, 2022), PVLDB (2015, 2019–2021), KDD (2019–2021), CHI (2021), IEEE VIS (2021), ICDE (2013, 2018), EDBT (2017), SDM (2017), CIKM (2011, 2012)

Journal Reviewer VLDB Journal (2009 – 2011, 2017, 2020 – 2021), TKDE (2007, 2011, 2012, 2016, 2018, 2020 – 2021), TKDD (2012), TODS (2013), TWEB (2012, 2015)

Selected Publications, Patents, and Grants

Traditional DP [1] Abdulhakim Qahtan, **Nan Tang**, Mourad Ouzzani, Yang Cao, and Michael Stonebraker. *Pattern Functional Dependencies for Data Cleaning.* PVLDB 2020.

[2] John K. Feser, Samuel Madden, **Nan Tang**, and Armando Solar-Lezama. *Deductive Optimization of Relational Data Storage.* OOPSLA 2020.

[3] El Kindi Rezig, Lei Cao, Giovanni Simonini, Maxime Schoemans, Samuel Madden, Mourad Ouzzani, **Nan Tang**, and Michael Stonebraker. *Dagger: A Data (not code) Debugger.* CIDR 2020.

[4] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and **Nan Tang**. *Raha: A Configuration-Free Error Detection System.* SIGMOD, 2019.

[5] Dong Deng, Wenbo Tao, Ziawasch Abedjan, Ahmed Elmagarmid, Ihab F. Ilyas, Guoliang Li, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and **Nan Tang**. *Unsupervised String Transformation Learning for Entity Consolidation.* ICDE, 2019.

[6] Abdulhakim A. Qahtan, Ahmed Elmagarmid, Raul Castro Fernandez, Mourad Ouzzani, and **Nan Tang**. *FAHES: A Robust Disguised Missing Values Detector.* KDD, 2018.

[7] Rohit Singh, Vamsi Meduri, Ahmed Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Armando Solar-Lezama, and **Nan Tang**. *Synthesizing Entity Matching Rules by Examples.* PVLDB, 2018.

[8] Raul Castro Fernandez, Essam Mansour, Abdulhakim Qahtan, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and **Nan Tang**. *Seeping Semantics: Linking Datasets using Word Embeddings for Data Discovery.* ICDE, 2018.

[9] Shuang Hao, **Nan Tang**, Guoliang Li, and Jianhua Feng. *Discovering Mis-Categorized Entities.* ICDE, 2018.

[10] Saravanan Thirumuruganathan, Laure Berti-Equille, Mourad Ouzzani, Jorge-Arnulfo Quiane-Ruiz, and **Nan Tang**. *UGuide – User-Guided Discovery of FD-Detectable Errors.* SIGMOD, 2017.

[11] Shuang Hao, **Nan Tang**, Guoliang Li, Jian Li, and Jianhua Feng. *Cleaning Relations using Knowledge Bases.* ICDE, 2017.

[12] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and **Nan Tang**. *The Data Civilizer System.* CIDR, 2017.

[13] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and **Nan Tang**. *Interactive and Deterministic Data Cleaning: A Tossed Stone Raises a Thousand Ripples.* SIGMOD, 2016.

[14] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and **Nan Tang**. *Detecting Data Errors: Where are we and what needs to be done?* PVLDB, 2016.

[15] Zuhair Khayyat, Ihab F. Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, **Nan Tang**, and Si Yin. *BigDansing: A System for Big Data Cleansing.* SIGMOD, 2015.

[16] Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, **Nan Tang**, and Yin Ye. *KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing.* SIGMOD, 2015.

[17] Matteo Interlandi, and **Nan Tang**. *Proof Positive and Negative in Data Cleaning.* ICDE, 2015.

[18] Jiannan Wang, and **Nan Tang**. *Towards Dependable Data Repairing with Fixing Rules.* ICDE, 2015.

[19] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and **Nan Tang**. *NADEEF: A Commodity Data Cleaning System.* SIGMOD, 2013.

[20] Wenfei Fan, Floris Geerts, **Nan Tang**, and Wenyuan Yu. *Inferring Data Currency and Consistency for Conflict Resolution.* ICDE, 2013.

[21] Wenfei Fan, Jianzhong Li, **Nan Tang**, and Wenyuan Yu. *Incremental Detection of Inconsistencies in Distributed Data.* ICDE, 2012.

[22] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Interaction Between Record Matching and Data Repairing.* SIGMOD, 2011.

[23] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data.* PVLDB, 2010. (The best paper award)

DL for DP

[24] **Nan Tang**, Ju Fan, Fangyi Li, Jianhong Tu, Xiaoyong Du, Guoliang Li, Sam Madden, and Mourad Ouzzani. *RPT: Relational Pre-trained Transformer Is Almost All You Need for Democratizing Data Preparation.* PVLDB, 2021.

[25] Saravanan Thirumuruganathan, **Nan Tang**, Mourad Ouzzani, and AnHai Doan. *Data Curation with Deep Learning.* EDBT, 2020.

[26] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and **Nan Tang**. *Distributed Representations of Tuples for Entity Resolution.* PVLDB, 2018.

DP for ML

[27] Tongyu Liu, Yinqing Luo, Ju Fan, **Nan Tang**, Guoliang Li, and Xiaoyong Du. *Adaptive Data Augmentation for Supervised Learning over Missing Data.* PVLDB, 2021.

VIS for DP

[28] Yuyu Luo, **Nan Tang**, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. *Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks.* SIGMOD, 2021.

[29] Yuyu Luo, Chengliang Chai, Xuedi Qin, **Nan Tang**, and Guoliang Li. *Interactive Cleaning for Progressive Visualization through Composite Questions.* ICDE, 2020.

[30] Yuyu Luo, **Nan Tang**, Guoliang Li, Tianyu Zhao, Wenbo Li, and Xiang Yu. *DEEPEYE: A Data Science System for Monitoring and Exploring COVID-19 Data.* IEEE Data Engineering Bulletin, 2020.

[31] Xuedi Qin, Yuyu Luo, **Nan Tang**, and Guoliang Li *Making Data Visualization More Efficient and Effective: A Survey.* VLDBJ, 2020.

[32] Xuedi Qin, Yuyu Luo, **Nan Tang**, and Guoliang Li. *DeepEye: Visualizing Your Data by Keyword Search.* EDBT, 2018.

[33] Xuedi Qin, Yuyu Luo, **Nan Tang**, and Guoliang Li. *DeepEye: Towards Automatic Data Visualization.* ICDE, 2018.

Miscellaneous

[34] Ji Sun, Guoliang Li, and **Nan Tang**. *Learned Cardinality Estimation for Similarity Queries*. SIGMOD, 2021.

[35] Xiang Yu, Guoliang Li, Chengliang Chai, and **Nan Tang**. *Reinforcement Learning with Tree-LSTM for Join Order Selection*. ICDE, 2020.

[36] Yong Wang, Guoliang Li, and **Nan Tang**. *Querying Shortest Paths on Time Dependent Road Networks*. PVLDB, 2019.

[37] Mourad Ouzzani, **Nan Tang**, and Raul Castro Fernandez. *Data Civilizer: End-to-End Support for Data Discovery, Integration, And Cleaning [Book Chapter]*. Making Databases Work: The Pragmatic Wisdom of Michael Stonebraker, 2019.

[38] **Nan Tang**, Qing Chen, and Prasenjit Mitra. *Graph Stream Summarization: From Big Bang to Big Crunch*. SIGMOD, 2016.

[39] Zuhair Khayyat, William Lucia, Meghna Singh, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, **Nan Tang**, and Panos Kalnis. *Lightning Fast and Space Efficient Inequality Joins*. PVLDB, 2016.

[40] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Yinghui Wu. *Adding Regular Expressions to Graph Reachability and Pattern Queries*. ICDE, 2011.

[41] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, Yinghui Wu, and Yunpeng Wu. *Graph pattern Matching: From Intractable to Polynomial Time*. PVLDB, 2010.

[42] **Nan Tang**, Lefteris Sidirourgos, and Peter Boncz. *Space-Economical Q-Gram Index for Exact String Matching*. CIKM, 2009.

[43] Ying Zhang, **Nan Tang**, and Peter Boncz. *Efficient Distribution of Full-Fledged XQuery*. ICDE, 2009.

Invited Papers

[44] Zuhair Khayyat, William Lucia, Meghna Singh, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, **Nan Tang**, and Panos Kalnis. *Fast and Scalable Inequality Joins*. VLDBJ, 2017. (Special issue: Best Papers of VLDB 2015)

[45] Wenfei Fan, Floris Geerts, **Nan Tang**, and Wenyuan Yu. *Conflict Resolution with Data Currency and Consistency*. ACM Journal of Data and Information Quality (JDIQ), 2014.

[46] Wenfei Fan, Floris Geerts, **Nan Tang**, and Wenyuan Yu. *Incremental Detection of Inconsistencies in Distributed Data*. TKDE, 2014. (Special issue: Best Papers of ICDE 2012)

[47] Wenfei Fan, Floris Geerts, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Data Quality Problems beyond Consistency and Deduplication*. In search of elegance in the theory and practice of computation: a Festschrift in honour of Peter Buneman, Edinburgh, UK, 2013.

[48] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data*. VLDBJ, 2012. (Special issue: Best Papers of VLDB 2010)

[49] Ying Zhang, **Nan Tang**, and Peter Boncz. *Projective Distribution of Full-Fledged XQuery*. TKDE, 2010. (Special issue: Best Papers of ICDE 2009)

Tutorials

[50] **Nan Tang**, Eugene Wu, and Guoliang Li. *Towards Democratizing Relational Data Visualization*. SIGMOD tutorial, 2019.

Demos

[51] El Kindi Rezig, Ashrita Brahmaroutu, Nesime Tatbul, Mourad Ouzzani, **Nan Tang**, Timothy Mattson, Samuel Madden, and Michael Stonebraker. *Debugging Large-Scale Data Science Pipelines using Dagger*. VLDB demo, 2020.

[52] Yuyu Luo, Chengliang Chai, Xuedi Qin, **Nan Tang**, Guoliang Li. *VisClean: Interactive Cleaning for Progressive Visualization*. VLDB demo, 2020.

[53] Yuyu Luo, Wenbo Li, Tianyu Zhao, Xiang Yu, Lixi Zhang, Guoliang Li, and **Nan Tang**. *DeepTrack: Monitoring and Exploring Spatio-Temporal Data (A Case of Tracking COVID-19)*. VLDB demo, 2020.

[54] Mashaal Musleh, Mourad Ouzzani, **Nan Tang**, and AnHai Doan. *CoClean: Collaborative Data Cleaning*. SIGMOD demo, 2020.

[55] El Kindi Rezig, Lei Cao, Michael Stonebraker, Giovanni Simonin, Wenbo Tao, Samuel Madden, Mourad Ouzzani, **Nan Tang**, and Ahmed K. Elmagarmid. *Data Civilizer 2.0: A Holistic Framework for Data Preparation and Analytics*. VLDB demo, 2019.

[56] Abdulhakim Qahtan, **Nan Tang**, Mourad Ouzzani, Yang Cao, and Michael Stonebraker. *ANMAT: Automatic Knowledge Discovery and Error Detection through Pattern Functional Dependencies*. SIGMOD demo, 2019.

[57] Essam Mansour, Dong Deng, Raul Castro Fernandez, Abdulhakim Qahtan, Wenbo Tao, Ziawasch Abedjan, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and **Nan Tang**. *Building Data Civilizer Pipelines with an Advanced Workflow Engine*. ICDE demo, 2018.

Patents

[58] *Dependable Data Repairing with Fixing Rules*. QCRI, HBKU (PCT/EP2013/052476).

[59] *Towards Dependable Data Repairing with Fixing Rules*. QCRI, HBKU (PCT/EP2014/052494).

[60] *KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing*. QCRI, HBKU (PCT/GB2014/051670).

[61] *NADEEF: A Holistic and Extensible Data Cleaning Platform*. QCRI, HBKU (PCT/EP2012/062446).

[62] *Generalized Data Cleaning using SAT-Solvers*. QCRI, HBKU (PCT/EP2012/062445).

Grants

[63] *Credible Open Knowledge Network* (NSF grant #1937143). Start date: September 1, 2019. End Date: May 31, 2021. Prof. Chengkai Li from the University of Texas at Arlington is the PI and I serve as a strategic partner.

[64] *Effective and Efficient Data Quality Management for Data Lakes* (Australian Research Council: DP210103593). From 2021 to present. Professor Wei Wang from the University of New South Wales is the PI and I serve as a co-PI.

Invited Talks

2020/05 *Data Visualization and Exploration of COVID-19 data*, QCRI lectures on the use of AI techniques for COVID-19, Qatar.

2019/10 *Data Preparation meets Data Visualization*, at Northeastern University, US.

2016/10 *Mind Your Analytics, Clean Your Data*, at Harvard University, US.

2016/03 *Graph Stream Summarization*, at MIT, US.

2015/12 *Trusted Data Cleaning*, at KAUST, Saudi Arabia.

2014/09 *Big Data Cleaning*, at Asia-Pacific Web Conference 2014, Distinguished Lecturer series.